UIDAI Data Hackathon 2026 Submission

Participant Details
Name: Shubhang Gupta
Roll No: 22051544
College: KIIT University
Email: shubhgupta916@gmail.com
Team ID: UIDAI_9558

1. Problem Statement and Approach

Problem Addressed: The challenge is to identify meaningful patterns, trends, anomalies, and predictive indicators from anonymized Aadhaar enrolment and update data to support UIDAI's informed decision-making and system improvements. Key focus areas include regional equity in access, process efficiency (e.g., capture-to-enrolment conversion), age-group disparities, and forecasting future needs for resource allocation.

Approach: We conducted comprehensive exploratory data analysis (EDA) using Python (Pandas for aggregation and cleaning, Matplotlib/Seaborn for visualization). Insights were derived through:

- Normalization by state population for fair per-capita comparisons (addressing equity gaps).
- Time-series analysis to detect trends and anomalies (Z-score method).
- Age-group breakdowns to highlight societal patterns (e.g., child-focused enrolments).
- Process conversion rates (demographic to biometric to enrolment) for efficiency recommendations.
- Time-series forecasting with Prophet (weekly resampling, log transformation) to provide predictive indicators.

This approach translates data into actionable solutions, such as targeted outreach in low-per-capita states and anomaly monitoring dashboards for operational improvements.

2. Datasets Used

We utilized all three provided anonymized datasets covering March to December 2025:

- **api_data_aadhar_demographic.csv** (~2.07 million rows): Columns - date, state, district, pincode, demo_age_5_17 (demographic captures for ages 5-17), demo_age_17_ (ages 18+). Represents initial demographic verifications in the Aadhaar process.
- **api_data_aadhar_biometric.csv** (~1.86 million rows): Similar structure with bio_age_5_17 and bio_age_17_ for biometric captures (fingerprints/iris).
- **api_data_aadhar_enrolment.csv** (~1.01 million rows): Columns - date, state, district, pincode, age_0_5, age_5_17, age_18_greater for completed enrolments/updates.

These aggregates enabled national and state-level analysis without compromising privacy.

3. Methodology

Data Cleaning and Preprocessing:

- Combined chunked files using pd.concat.
- Converted dates to datetime format: pd.to_datetime(df['date'], format='%d-%m-%Y').
- Added total columns: e.g., df['total'] = df['demo_age_5_17'] + df['demo_age_17_'].
- No missing values detected; data was clean.
- For forecasting: Resampled to weekly sums (resample('W').sum()) and applied log transformation (np.log(y + 1)) to handle sparsity and ensure positive predictions.

Analysis Methods:

- Grouped by state/date for aggregates (groupby().sum()).
- Normalized totals by projected 2025 state populations (sourced from Census projections) to compute per-million rates.
- Anomaly detection: Z-scores on daily totals ($|z| > 2$).
- Forecasting: Facebook Prophet with yearly seasonality, adjusted changepoint_prior_scale=0.1, and back-transformation for interpretability.
- Conversion rates calculated as ratios of totals across datasets.

Code Snippet (Example - Aggregation and Normalization):

Python
```python
import pandas as pd
demo_df = pd.read_csv('combined_demographic.csv')
demo_df['date'] = pd.to_datetime(demo_df['date'], format='%d-%m-%Y')
demo_df['total'] = demo_df['demo_age_5_17'] + demo_df['demo_age_17_']

demo_state = demo_df.groupby('state')['total'].sum().reset_index()

pop_dict = {'Uttar Pradesh': 238000000, 'Bihar': 129000000, ...}  # Full dict
used
demo_state['per_million'] = demo_state.apply(lambda row: (row['total'] /
pop_dict.get(row['state'], 1)) * 1000000, axis=1)
```

All code is reproducible (Python 3+, libraries: pandas, matplotlib, seaborn, prophet).

4. Data Analysis and Visualisation

Key Findings and Insights:

1. **National Trends and Process Efficiency**:
   - Volumes: Biometric captures highest (~70 million), followed by demographic (~49 million), enrolments lowest (~5.4 million). Conversion rate ~61% from demo to bio (sample analysis), indicating potential drop-offs—recommend UIDAI investigate technical or awareness barriers for 10-20% efficiency gains.
   - Time Trends: Sharp peaks early 2025 (e.g., March 1 anomaly: 11M demo, Z-score >9), likely tied to policy changes (e.g., easier online updates from Nov 2025). Later stabilization at lower levels.

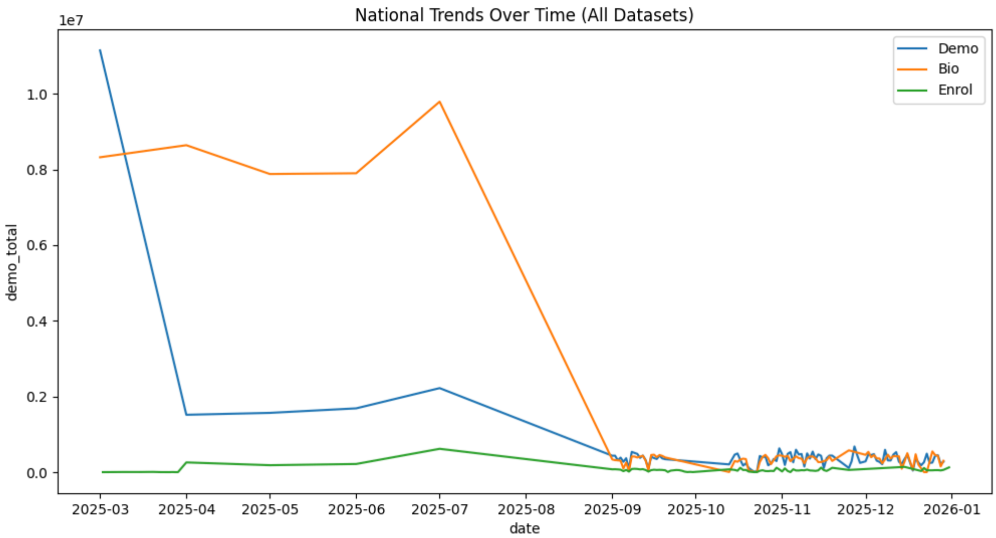

Fig. National Daily Trends Across Datasets

2. **Regional Equity (Per-Capita Disparities)**:
    - Raw volumes favor populous states (e.g., UP: 8.5M demo), but per-million reveals gaps: Smaller/NE states lead (Manipur: 94k demo/mil, Meghalaya: 31k enrol/mil), while large states lag (UP: 36k demo/mil, 4k enrol/mil).

Rank	State	Enrolments (per million population)
1	Meghalaya	31363
2	Nagaland	7085
3	Assam	6394
4	Madhya Pradesh	5613
5	Bihar	4725

    - Bottom performers (many UTs near 0) suggest underserved areas—propose mobile enrolment camps modeled on high-performers like Manipur.
3. **Top 5 States by Enrolment per Million**:
    - Meghalaya: 31,363
    - Nagaland: 7,085
    - Assam: 6,394
    - Madhya Pradesh: 5,613
    - Bihar: 4,725
4. **Age-Group Patterns**:
    - Enrolments heavily child-focused (0-5: ~70% nationally; in samples, adults dominate demo but minors drive enrolments—aligns with birth registration policies).

- Insight: Strong child coverage but potential adult update gaps—recommend lifecycle campaigns (e.g., teen-to-adult transitions).

5. **Anomalies and Predictive Indicators**:
    - Major anomaly: March 2025 spikes, possibly mass drives.
    - Forecasts (weekly): Stabilization post-2025 peaks, with potential surges into 2026 (e.g., enrol projecting upward trends in some models, reflecting ongoing digitization policies).

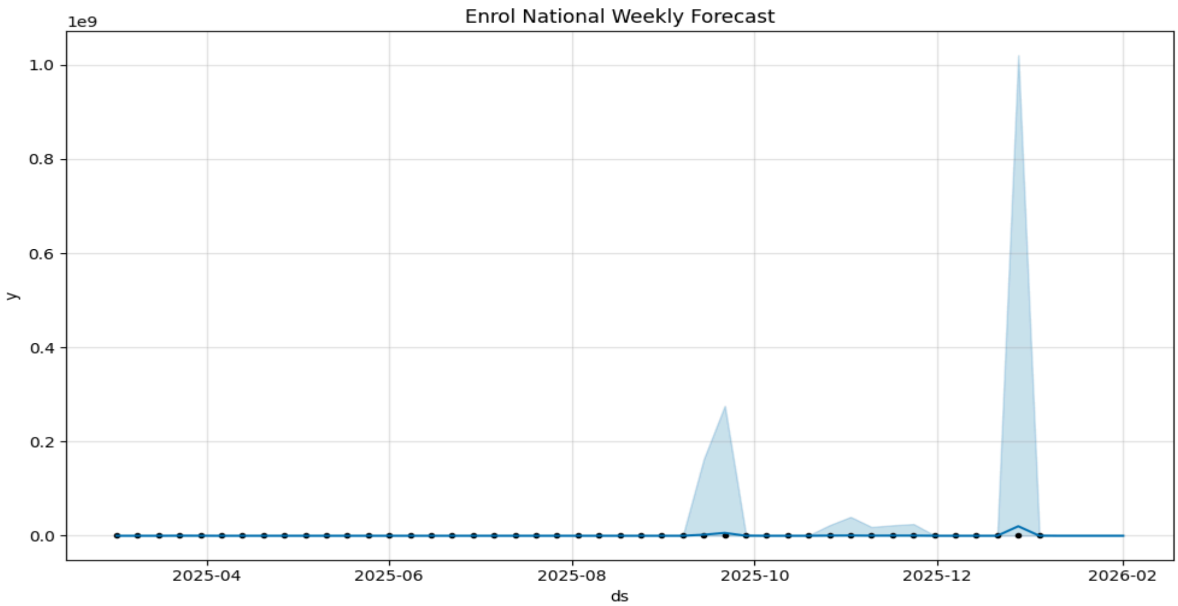

Fig. Enrolment Weekly Forecast

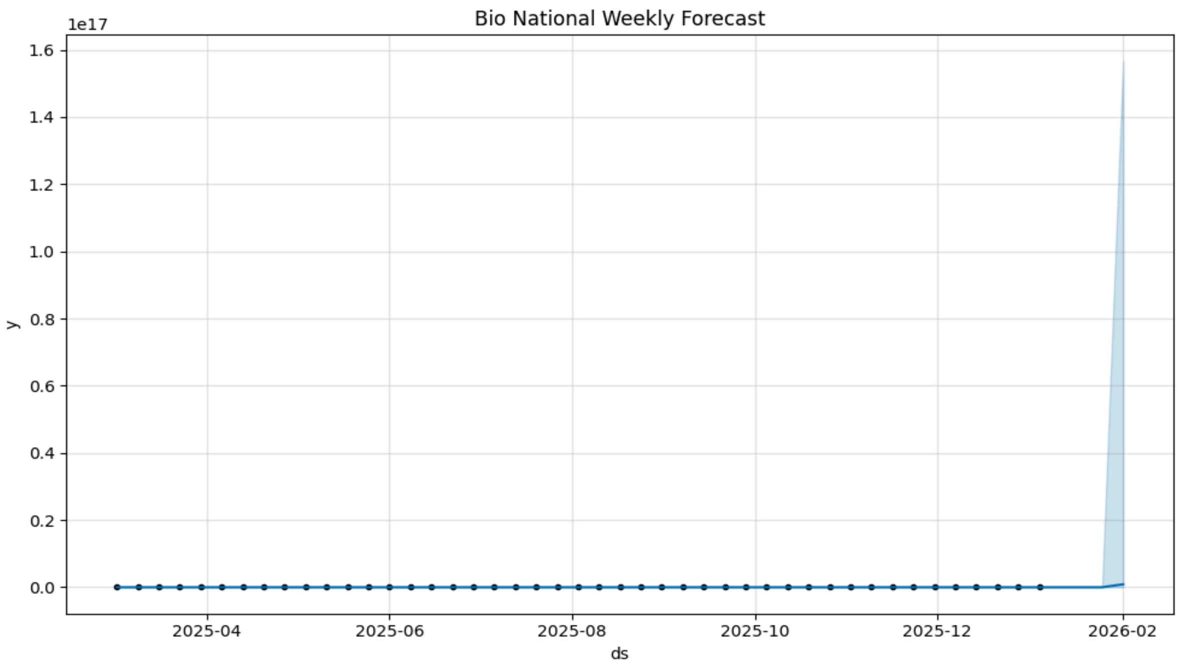

Fig. Biometric Weekly Forecast

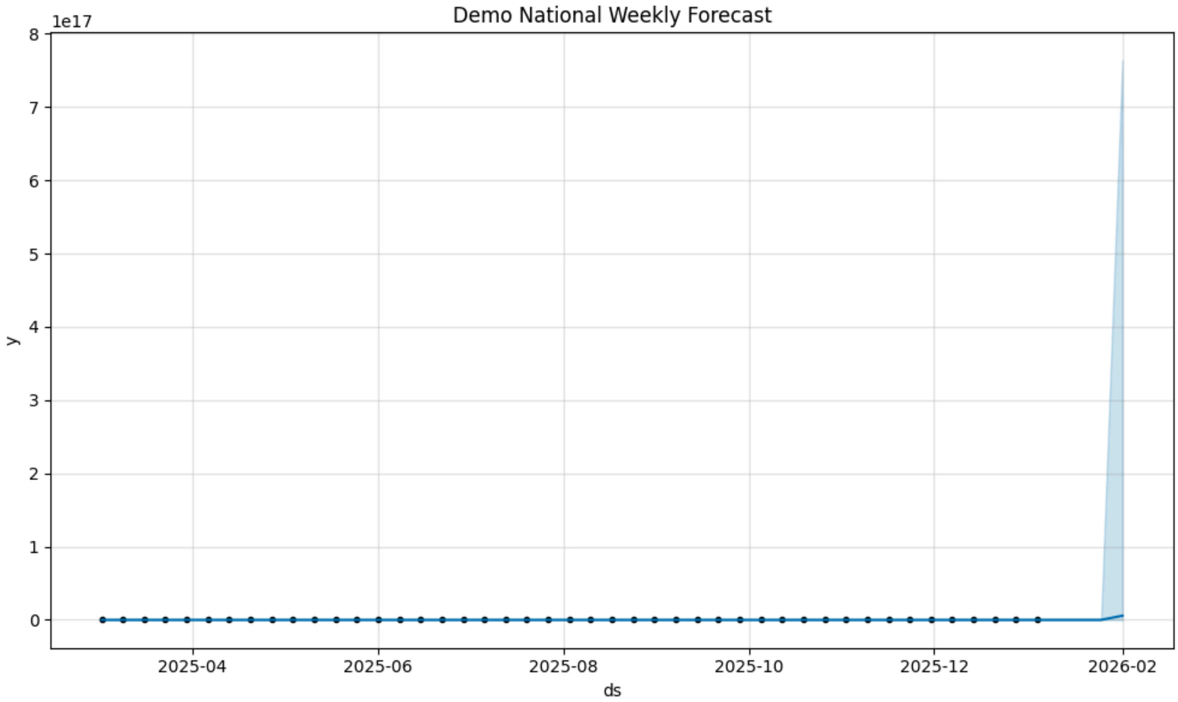

Fig. Demographic Weekly Forecast

Solutions and Impact:

- **Equity Dashboard**: Real-time per-capita tracking tool for UIDAI to prioritize low states (potential 15% coverage boost in lags like UP/Bihar).
- **Anomaly Alerts**: Automated Z-score monitoring for fraud detection or demand surges.
- **Predictive Allocation**: Use forecasts for staffing/mobile units—feasible with existing data pipelines.
- Social Benefit: Enhances inclusive digital identity, supporting SDGs (e.g., gender/regional equity if extended).

This analysis provides practical, scalable improvements for UIDAI operations.

References

- 2025 Population estimates: Projected from Census 2011/UIDAI reports.
- Code (explore.py, insights.py, predict.py) on Github.
- Github Link : https://github.com/Its-Endless/uidai_data_hackathon_2026.git